Modulation of Coffee Flavour Precursor Levels in Green Coffee Grains

Background of the Invention:

Coffee contains a highly complex mixture of flavour molecules. Extensive research on the composition of instant and fresh ground coffee beverages has, to date, identified more than 850 compounds, many of which are flavour active molecules (Flament, I (2002) Coffee Flavor Chemistry, John Wiley and Sons, UK). However, few of the final coffee flavour molecules found in the cup of coffee are present in the raw material, the green grain (green beans) of the plant species *Coffea arabica* or *Coffea canephora (robusta)*. In fact, the majority of the coffee flavour compounds are generated during one or more of the multiple processing steps that occur from the harvest of the ripe red coffee cherries to the final roasted ground coffee product, or extracts thereof, for example soluble coffee products.

The various steps in the production of coffee are described in Smith, A.W., in Coffee; Volume 1: Chemistry pp 1-41, Clark, R.J. and Macrea, R. eds, Elsevier Applied Science London and New York, 1985; Clarke, R.J., in Coffee: Botany, Biochemistry, and Production of Beans and Beverage, pp 230-250 and pp 375-393; and Clifford, M.N. and Willson, K.C. eds, Croom Helm Ltd, London. Briefly, the process starts with the collection of mature, ripe red cherries. The outer layer, or pericarp, can then be removed using either the dry or wet process. The dry process is the simplest and involves 1) classification and washing of the cherries, 2) drying the cherries after grading (either air drying or mechanical drying), and 3) dehusking the dried cherries to remove the dried pericarp. The wet process is slightly more complicated, and generally leads to the production of higher quality green beans. The wet process is more often associated with *C. arabica* cherries. The wet process consists of 1) classification of the cherries, 2) pulping of the cherries, this step is done soon after harvest and generally involves mechanical removal of the "pulp", or pericarp, of the mature cherries, 3) "fermentation", the mucilage that remains attached to the grain of the cherries after pulping is removed by allowing the grain plus attached mucilage to be incubated with water in tanks using a batch process. The "fermentation" process is allowed to continue up to 80 hours, although often 24 hours is generally enough to allow an acceptable fermentation and to cause the pH to drop from around 6.8-6.9 to 4.2-4.6, due to various enzymatic activities and the metabolic action of microorganisms which grow during the fermentation, 4) drying, this step involves either air or mechanical hot air drying of the fermented coffee grain and 5) "hulling", this step involves the mechanical removal of the "parch" of the dried coffee grain (dried parchment coffee) and often the silverskin is also removed at this stage. After wet or dry processing, the resulting green coffee grain are often sorted, with most sorting procedures being based on grain size and/or shape.

The next step in coffee processing is the roasting of the green grain after dehusking or dehulling of dry or wet processed coffee, respectively. This is a time-dependent process which induces significant chemical changes in the bean. The first phase of roasting occurs when the supplied heat drives out the remaining water in the grain. When the bulk of the water is gone, roasting proper starts as the temperature rises towards 190-200°C. The degree of roasting, which is usually monitored by the colour development of the beans, plays a major role in determining the flavour characteristics of the final beverage product. Thus, the time and temperature of the roasting are tightly controlled in order to achieve the desired coffee flavour profile. After roasting, the coffee is ground to facilitate extraction during the production of the coffee beverage or coffee extracts (the latter to be used to produce instant coffee products). Again, the type of grinding can influence the final flavour of the beverage.

While a considerable amount of research has been carried out on the identification of the flavour molecules in coffee, much less work has been done regarding the physical and chemical reactions which occur within the coffee grains during each of the processing steps. This latter point is particularly evident for the roasting reaction, where the large number of grain constituents undergo an extremely complex series of heat induced reactions (Homma, S. 2001, In "Coffee: Recent Developments". R.J. Clarke and O.G. Vitzthum eds, Blackwell Science, London; Yeretzian, C., *et al* ((2002) Eur. Food Res. Technol. 214, 92-104; Flament, I (2002) Coffee Flavor Chemistry, John Wiley and Sons, UK; Reineccius, G.A., "The Maillard Reaction and Coffee Flavor" Conference Proceedings of ASIC, 16th Colloque, Kyoto, Japan 1995).

While the details of most of the reactions that occur during the different steps of coffee processing remain relatively unclear, it is thought that an important flavour generating reaction responsible for many of the flavours associated with coffee aroma is the "Maillard" reaction during coffee roasting. A vigorous Maillard reaction occurs between the grain reducing sugars/polysaccharide degradation products and the amino group containing molecules (particularly the proteins, peptides, and amino acids) during the roasting step.

Because the Maillard reaction apparently makes an important contribution to the generation of coffee flavour and aroma molecules during coffee roasting, there might be an association between the levels of primary Maillard reactants in the green beans and the quality of the flavour/aroma developed after roasting.

As noted above, an important group of substrates in the Maillard reaction are amino acids, peptides and proteins. Using 2-D electrophoresis, it has been shown that differences exist in the levels and amounts of the major storage proteins in *arabica* and *robusta* green coffee beans – however, no association between these storage protein differences and flavour quality was noted (Rogers *et al*, 1999, Plant Physiol. Biochem. Vol 37, 261-272). It has also recently been found that small differences exist between the storage proteins of immature and mature coffee beans, which have different flavour qualities (Montavon, P. *et al*, 2003, J. Agric and Food Chemistry Vol 51, 2328-2334). Because there are many changes occurring during seed maturation, this latter work suggests a link may exist between the quality improvement caused by maturation and the differences seen in the 2-D gel patterns of the main coffee storage proteins.

It has recently been shown that there are differences in the profiles of peptides isolated from *arabica* and *robusta* green beans (Ludwig *et al* 2000, Eur. Food Res Technol., Vol 211, 111-116.). Although their results showed that the *arabica* and *robusta* peptide extracts differ in their aroma precursor profile, the data presented in this report do not identify which component(s) in the extracts is / are responsible for these aroma profile differences. These workers also detected at least two different proteinase activities in crude extracts of the green coffee, but they did not correlate any specific activities with aroma/flavour quality (Ludwig *et al* 2000, Eur. Food Res Technol., Vol 211, 111-116). Finally, it is also thought that the very high temperatures used during the later stages of green coffee grain roasting cause substantial cleavage of the proteins present in the coffee grain (Homma, S. 2001, In "Coffee: Recent Developments". R.J. Clarke and O.G. Vitzthum eds, Blackwell Science, London; Montavon, P., *et al* 2003, "Changes in green coffee protein profiles during roasting", J. Agric. Food Chem. 51, 2335-2343). However, the overall scheme for this protein degradation is very poorly understood, but presumably depends on, among other things, the precise state of the main coffee proteins in the raw material before the start of roasting. To our knowledge, there are no other significant reports addressing the possibility that peptide profiles in coffee could be involved in the production of coffee aroma/flavour.

In the roasting of the fermented seeds of *Theobroma cacao* (cocoa beans), there would appear to be an involvement of seed amino acids and peptides in the development of Maillard reaction aromas/flavours. Relative to other seeds, *T. cacao* seeds have been shown to have an unusually high level of aspartic proteinase activity (Biehl, B., Voigt, J., Voigt, G., Heinrichs, H., Senyuk, V. and Bytof, G. (1994) "pH dependent enzymatic formation of oligopeptides and amino acids, the aroma precursors in raw cocoa beans". In *The Proceedings of the 11th International Cocoa Research Conference*, 18-24 July 1993, Yamoussoukro, Ivory Coast). In order to produce cocoa beans with a high level of cocoa flavour precursors, it is necessary to carry out a natural fermentation step (unfermented beans develop little flavour when roasted). During this fermentation step, the sugars in the pulp are fermented, generating high levels of acids, particularly acetic acid (Carr, J.G. (1982) Cocoa. In *Fermented Foods*. Economic Microbiology. Vol 7. pages 275-292. (A.H. Rose ed). Academic Press). As the fermentation continues, the pH in the seed decreases and the cell structure becomes disrupted. The low pH triggers the abundant cacao seed aspartic proteinase to become mobilized and/or activated, resulting in a massive degradation of cellular protein (Biehl, B., Passern, D., and Sagemann, W. (1982) "Effect of Acetic Acid on Subcellular Structures of Cocoa Bean Cotylydons". J. Sci. Food Agric. 33, 1101-1109; Biehl., B., Brunner, E., Passern, D., Quesnel, V.C., and Adomako, D. (1985) "Acidification, proteolysis and flavour potential in fermenting cocoa beans". J. Sci. Food Agric. 36, 583-598). Peptides and amino acids have been shown to be cocoa flavour precursors (Rohan, T. (1964) "The precursors of chocolate aroma: a comparative study of fermented and unfermented cocoa beans". J. Food Sci., 29, 456-459; Voigt, J. and Biehl, B. (1995) "Precursors of the cocoa specific aroma components are derived from the vicilin-class (7S) globulin of the cocoa seeds by proteolytic processing". Bot. Acta 108, 283-289). Thus, the *T. cacao* seed aspartic proteinase, together with a seed serine carboxypeptidase, have been proposed to be critical for the generation of cocoa flavour precursors during fermentation (Voigt, J. and Biehl, B. (1995) "Precursors of the cocoa specific aroma components are derived from the vicilin-class (7S) globulin of the cocoa seeds by proteolytic processing". Bot. Acta 108, 283-289; Voigt, J., Heinrichs, H., Voigt, G. and Biehl, B. (1994) "Cocoa-specific aroma precursors are generated by proteolytic digestion of the vicilin-like globulin of cocoa seeds". Food Chemistry, 50, 177-184.) The gene encoding the abundant cacao seed aspartic proteinase has been identified and a method to over-express this protein in cacao seeds which can generate increased levels of cacao flavour precursor amino acids and peptides in fermented cocoa beans has recently been described in International Patent Publication No. 02/04617, the whole contents of which are incorporated herein by reference. However, the teaching of International Patent Publication No. 02/04617 is directed towards cacao seeds, which undergo a specific long acid fermentation step, unlike coffee grains which do not.

An important vacuolar cysteine proteinase (CP) is the KDEL containing cysteine proteinase. This type of proteinase has been characterized in several plants. To date, three genes encoding cysteine proteinases with C-terminal KDEL sequences have been found in arabidopsis (Gietl, C., and Schmid, M. 2001, Naturwissenschaften **88**, 49-58). One is expressed in senescing ovules, one in vascular vessels, and the third in maturing siliques. However, more detailed studies on this protein have been done in other plants. For example, a CP called the sulfhydryl-endoproteinase (SH-EP) has been characterized in the cotyledons of *Vigna mungo* seeds (Toyooka, K., Okamoto, T., and Minamikawa, T. (200) J. Cell Biol. 148, 453-463.). SH-EP is expressed *de-novo* in germinating cotyledons of *V. mungo*, and is proposed to be involved in the degradation of storage proteins accumulated in the protein storage vacuoles (Okamato, T and Minamikawa, T. J. Plant Physiol. **152**, 675-682). A key feature of the SH-EP polypeptide is that it possesses a specific COOH terminal sequence KDEL which directs the transport of this protein from the endoplasmic reticulum (ER) to the protein storage vacuoles (Toyooka et al., 2000). It has also been recently proposed that the SH-EP protein is actually involved, via the presence of its KDEL sequence, in the formation of specific vesicles called KV (KDEL Vesicles) in a previously undescribed vesicle transport system (Okamato, T., Shimada, T., Hara-Nishimura, I., Nishimura, M., and Minamikawa, T. (2003) Plant Physiology, 132, 1892-1900).

A related proposal has been made for a KDEL containing CP protein found in germinating castor bean cotyledons (*Ricinus communis*). In this plant, the authors implicate this KDEL proteinase in the programmed cell death of the endosperm to continue suppling nutrients for the germinating castor bean embryo (Gietl, C., and Schmid, M. 2001, Naturwissenschaften 88, 49-58). These authors propose that, in the castor bean, the KDEL proteinase is made in the ER of germinating seeds before day 3. When the seed coat is cast off, around day 3, the KDEL containing CP then gets packaged into a specific vesicle called a ricinosome. Later, as the endosperm becomes soft between day 4-5, the KDEL-CP has its anchor sequence (KDEL) cleaved off and this proteinase migrates to the cytoplasm where it assists in the general degradation of the cellular protein.

Objects of the Invention

It is an object of the present invention to modify protein/peptide/amino acid flavour precursor pools in coffee.

More specifically, it is an object of the present invention to modify the levels of the flavour precursors in the raw material (the green grain) so that, following post harvest treatment and roast-processing, an altered flavour may be achieved. Without being bound by theory, it is believed that, if there are variations in the levels of peptides and protein degradation between coffees with significantly different flavours, then these variations could be due to differences in the endogenous proteinase activities in these different grains. This difference might be detectable at the level of mRNA expression by variations in the levels of expression for particular seed proteinase genes.

Statements of the Invention

The present invention involves, therefore, identifying gene sequences encoding for coffee grain (seed) specific proteinases and showing that there are indeed variations in the expression of these genes in *arabica* and *robusta.*

More specifically, the present invention discloses two major coffee cysteine proteinases (CcCP-1 and CcCP-4), four major coffee cysteine proteinase inhibitors (CcCPI-1, CcCPI-2, CcCPI-3 and CcCPI-4) and two coffee aspartic proteinases (CcAP-1 and CcAP-2), all of which are expressed in coffee seeds. We further show how either over-expression of these proteins specifically late in seed development, or the reduced expression of these proteins during late seed development, can alter the amino acid/peptide/protein profile of the mature beans. By using one or more of the disclosed gene sequences and gene constructs to alter the amino acid/peptide/protein profile of the mature beans, we disclose a new method to alter the flavour precursor profile of mature coffee beans.

In a first aspect, the present invention provides an isolated polynucleotide comprising a nucleotide sequence encoding a polypeptide having cysteine proteinase activity, wherein the amino acid sequence of the polypeptide and the amino acid sequence selected from SEQ ID Nos. 2 or 16 have at least 70%, preferably at least 80%, sequence identity based on the ClustalW alignment method; or the complement of the nucleotide sequence, wherein the complement contains the same number of nucleotides as the nucleotide sequence, and the complement and the nucleotide sequence are 100% complementary. Preferably, the amino acid sequence of the polypeptide and the amino acid sequence of SEQ ID Nos. 2 or 16 have at least 85%, preferably at least 90%, optionally at least 95%, sequence identity based on the ClustalW alignment method. Preferably, the nucleotide sequence comprises the nucleotide sequence of SEQ ID Nos. 1 or 15. Preferably, the polypeptide comprises the amino acid sequence of SEQ ID Nos. 2 or 16.

In a second aspect, there is provided an isolated polynucleotide comprising a nucleotide sequence encoding a polypeptide having cysteine proteinase inhibitor activity, wherein the amino acid sequence of the polypeptide and the amino acid sequence selected from SEQ ID Nos. 4, 10, 12 and 14 have at least 70%, preferably at least 80%, sequence identity based on the ClustalW alignment method; or the complement of the nucleotide sequence, wherein the complement contains the same number of nucleotides as the nucleotide sequence, and the complement and the nucleotide sequence are 100% complementary. Preferably, the amino acid sequence of the polypeptide and the amino acid sequence selected from SEQ ID Nos. 4, 10, 12 and 14 have at least 85%, preferably at least 90%, optionally at least 95%, sequence identity based on the ClustalW alignment method. Preferably, the nucleotide sequence comprises the nucleotide sequence selected from SEQ ID Nos. 3, 9, 11 or 13, optionally from SEQ ID Nos. 9, 11 or 13, further optionally from SEQ ID Nos. 9 or 13; still further optionally being SEQ ID No. 9. Preferably, the polypeptide comprises the amino acid sequence selected from SEQ ID Nos. 4, 10, 12 and 14, optionally from SEQ ID Nos. 10, 12 and 14, further optionally from SEQ ID Nos. 10 or 14; still further optionally being SEQ ID No. 10.

In a third aspect, there is provided an isolated polynucleotide comprising a nucleotide sequence encoding a polypeptide having aspartic endoproteinase activity, wherein the amino acid sequence of the polypeptide and the amino acid sequence selected from SEQ ID No. 6 or 8, preferably SEQ ID No. 8, have at least 75%, preferably at least 80%, sequence identity based on the ClustalW alignment method, or the complement of the nucleotide sequence, wherein the complement contains the same number of nucleotides as the nucleotide sequence, and the complement and the nucleotide sequence are 100% complementary. Preferably, the amino acid sequence of the polypeptide and the amino acid sequence selected from SEQ ID No. 6 or 8, preferably SEQ ID No. 8, have at least 85%, preferably at least 90%, optionally at least 95%, sequence identity based on the ClustalW alignment method. Preferably, the nucleotide sequence comprises the nucleotide sequence of SEQ ID No. 5 or 7, preferably SEQ ID No. 7. Preferably, the polypeptide comprises the amino acid sequence of SEQ ID No. 6 or 8, preferably SEQ ID No. 8.

In a further aspect, there is provided a vector comprising the polynucleotide of any one of first to third aspects of the invention.

In a further aspect, there is provided a non-native recombinant DNA construct comprising the polynucleotide of any one of first to third aspects of the invention, operably linked to a regulatory sequence. It will be appreciated that, in the non-native construct, either the polynucleotide is non-native or the regulatory sequence is non-native or both are non-native.

In a further aspect, there is provided a method for transforming a cell comprising transforming the cell with the polynucleotide of any one of first to third aspects of the present invention.

In a further aspect, there is provided a cell comprising the aforementioned non-native recombinant DNA construct, which cell is preferably a prokaryotic cell, an eukaryotic cell or a plant cell, preferably a coffee cell.

In a further aspect, there is provided a transgenic plant comprising such a transformed cell.

In the present application, coffee cherry terms are defined as follows: coffee cherry; entire fruit; exocarp, skin; pericarp, fleshy major outer layer of cherry; and grain, coffee seed. For a fuller explanation of these terms, reference is made to Clarke, R.J., in Coffee: Botany, Biochemistry, and Production of Beans and Beverage, pp 230, Clifford, M.N. and Willson, K.C. eds, Croom Helm Ltd, London, the contents of which are incorporated in their entirety.

Brief Description of the Invention

The invention can be understood from the following detailed description and the accompanying Sequence Listing which forms part of the present application.

Table 1 hereunder lists the polypeptides that are described herein, along with the corresponding sequence identifier (SEQ ID No) as used in the attached listing.

Table 1:

SEQ ID No 1 (CcCP1 : Cysteine proteinase, nucleic acid and its corresponding amino acid)
SEQ ID No 2 (CcCP1: Cysteine proteinase, amino acid)
SEQ ID No 3 (CcCPI-1 : Cysteine proteinase Inhibitor, nucleic acid and its corresponding amino acid)
SEQ ID No 4 (CcCPI-1: Cysteine proteinase Inhibitor, amino acid)
SEQ ID No 5 (CcAP1 : Aspartic endoproteinase 1, nucleic acid and its corresponding amino acid)
SEQ ID No 6 (CcAP1 : Aspartic endoproteinase 1, amino acid)
SEQ ID No 7 (CcAP2 : Aspartic proteinase 2, nucleic acid and its corresponding amino acid)
SEQ ID No 8 (CcAP2 : Aspartic proteinase 2, amino acid)
SEQ ID No 9 (CcCPI-2 : Cysteine proteinase Inhibitor, nucleic acid and its corresponding amino acid)
SEQ ID No 10(CcCPI-2 : Cysteine proteinase Inhibitor, amino acid)
SEQ ID No 11(CcCPI-3 : Cysteine proteinase Inhibitor, nucleic acid and its corresponding amino acid)
SEQ ID No 12(CcCPI-3 : Cysteine proteinase Inhibitor, amino acid)
SEQ ID No 13(CcCPI-4 : Cysteine proteinase Inhibitor, nucleic acid and its corresponding amino acid)
SEQ ID No 14 (CcCPI-4 : Cysteine proteinase Inhibitor, amino acid)
SEQ ID No 15(CcCP-4 : Cysteine proteinase, nucleic acid and its corresponding amino acid)
SEQ ID No 16(CcCP-4 : Cysteine proteinase, amino acid)

The sequence listing employs the one letter codes for nucleotide sequence characters and the three letter codes for amino acids as defined for IUPAC-IUBMB Standards and as described in Nucleic Acids Research 13:3021-3030 (1985), which is incorporated herein by reference.

Drawings

In the drawings,

Figure 1 shows a Northern blot analysis of cysteine proteinase gene in different tissues of *Coffea arabica*, in which the lanes are labeled R: root, S: stem, L: young leaves; and SG, LG, Y and Red are grain from small green fruit, large green fruit, yellow fruit and red fruit, respectively. Five micrograms of total RNA was loaded in each lane. MW is an RNA size ladder. Panel B illustrates an autoradiogram after 24 hours exposure showing the appearance of CcCP-1 mRNA in the tissues tested and Panel A demonstrates the ethidium bromide staining of the gels prior to blotting;

Figure 2 shows a Northern blot analysis of the expression of the Cysteine proteinase CcCP-1 gene in different tissues of *Coffea arabica*, in which lanes are labelled R, root; S, stem; L, young leaves; F, flowers. SG (G), LG (G), Y (G) and Red (G) correspond to RNA isolated from the grain of small green, large green, yellow and red cherries, respectively, and lanes which are labelled SG (P), LG (P), Y (P) and Red (P) correspond to RNA isolated from the pericarp tissue of small green, large green, yellow and red cherries, respectively. Five micrograms of total RNA was loaded in each lane. Panel A demonstrates the ethidium bromide staining of the large ribosomal RNA prior to blotting as a loading control, Panel B is an autoradiogram showing the appearance of the CcCP-1 mRNA in the specific tissues tested;

Figure 2A: Alignment of the full sequence of the protein encoded by CcCP-1 cDNA with other full-length cysteine proteinases available in the NCBI database. This was done in Megalign by the CLUSTAL method in the MegAlign (DNASTAR). Shaded blocks indicate identical amino acids. Accession numbers of the EMBL database are given in parentheses. Arabidopsis thaliana (AY070063); Vicia sativa (Z99172); Glycine max GMCP3 (Z32795); Glycine max GmPM33 (AF167986); Phaseolus vulgaris Moldavain (Z99955); Solanum melongena (AF082181); Nicotiana tabacum (AJ242994); Lycopersicon esculentum (Z14028); Vicia faba (AY161277).

Figure 3 shows a Northern blot analysis of Cysteine proteinase inhibitor (CcCPI-1) gene in different tissues of *Coffea arabica*, in which the lanes are labeled R: root, S: stem, L: young leaves and SG, LG, Y and Red for grain from small green fruit, large green fruit, yellow fruit and red fruit, respectively. Five micrograms of total RNA was loaded in each lane. MW is an RNA size ladder. Panel B illustrates an autoradiogram after 24 hours exposure and panel A demonstrates the ethidium bromide staining of the gels prior to blotting;

Figure 4 shows a Northern blot analysis of Cysteine proteinase inhibitor (CcCPI-1) gene in different stages of development of *Coffea arabica* (ARA) and *Coffea robusta* (ROB) fruit. The lanes are labeled small green fruit (SG), large green fruit (LG), yellow fruit (Y) and red fruit (Red), respectively. Five micrograms of total RNA was loaded in each lane. MW is an RNA size ladder. Panel B illustrates an autoradiogram after 24 hours exposure showing the appearance of CcCPI-1 mRNA in the specific tissues tested. Panel A demonstrates the ethidium bromide staining of the gels prior to blotting;

Figure 5 shows RT-PCR analysis of the expression of CcCP-1 during *Coffea arabica* grain germination. PCR reaction was carried out using 10 µl of each cDNA diluted 1/100. The cycling conditions were 2 min at 94°C, 35 cycles of 94°C, 61°C for 1.5 min, and 72°C for 2.5 min. The final extension step was for 7 min at 72°C. The PCR primers were:

A4-43-upper : 5'- ACCGAGGAGGAGTTTGAGGCTACG - 3'
A4-43-lower : 5'- ACGCTTCCCCCATGAGTTCTTGA - 3'

mRNAs were amplified by RT-PCR using specific primers (CcCP-1 up/CcCP-1 low) on different templates: cDNAs from sterilized seed (T0) and seeds taken after 2 days (2d), 3 days (3d), 5 days (5d), 1 month (1m) and 2 months (2m) of germination, respectively. The PCR products were resolved in a 1% (w/v) agarose gel and stained with ethidium bromide. RPL39; amplified fragment of cDNA encoding the L39 protein of the 60S ribosomal large subunit;

Figure 6 shows Western-blot analysis of the expression of CcCP1 protein (A). Total proteins were extracted from grains (g) and pericarp (p) collected from developing coffee cherries at stages Small Green (SG), Large Green (LG), Yellow (Y) and Red. Panel B -Separation of 50µg of total protein on a 12% SDS-PAGE gel and stained with Comassie blue. Panel A -

Protein detection was performed using a anti-CRP4 polyclonal antibody (rabbit) as described in the methods. Approximate size of bands in panel B are indicated with arrows at left. The large arrow inside each panel indicates the presence of a major storage protein that cross reacts with one of the antibodies;

Figure 6A shows the optimal alignment of the complete protein encoded by CcCPI-1 cDNA with other homologous full-length cysteine proteinases available in the NCBI. Shaded blocks indicate identical amino acids. Accession numbers of the EMBL database and percentage identities are given in parentheses. *Malus x domestica* (AAO18638; 42.3% identity), *Common sunflower* (JE0308; 41.5% identity), *Arabidopsis thaliana* (AAM64985; 30% identity) and *Rumex obtusifolius* (CAD21441; 29.3% identity);

Figure 7 shows RT-PCR analysis of the expression of CcCPI-1 gene in different tissues of *Coffea arabica* CCCA2 (A) *and Coffea robusta* FRT-32 (B). PCR reaction was carried out using 10 µl of each cDNA diluted 1/1000. The cycling conditions were 2 min at 94°C, 40 cycles of 94°C for 1 min, 60°C for 1.5 min, and 72°C for 1min. The final extension step was for 7 min at 72°C. The PCR primers were:

CcCPI-1 (up) 5' AGGAAAGTGGGAGCAAGGGAGAAGA 3'
CcCPI-1 (low) 5' TAGTATGAACCCAAGGCCGAACCAC 3'.

The lanes are labeled as follows: - M, Markers; +P, diluted plasmid containing the CcCPI-1 gene; R, root; S, stem; L, young leaves; F, flowers. - SG (G), LG (G), Y (G) and Red (G) are grain isolated from small green, large green, yellow and red cherries, respectively. SG (P), LG (P), Y (P) and Red (P) are pericarp tissue isolated from small green, large green, yellow and red cherries, respectively;

Figure 8 shows the optimal alignment of the complete protein encoded by CcCPI-2 cDNA with other homologous full-length cysteine proteinases available in the NCBI. Shaded blocks indicate identical amino acids. Accession numbers of the EMBL database and percentage identities are given in parentheses. *Rumex obtusifolius* (CAD21441; 66.7% identity), *Dianthus caryophyllus* (AAK30004; 71.7% identity), *Manihot esculenta* (AAF72202; 65.2% identity);

Figure 9 shows the RT-PCR analysis of the expression of CcCPI-2 gene in different tissues of *Coffea arabica* CCCA2 (A) *and Coffea robusta* FRT-32 (B). PCR reaction was carried out using 10 µl of each cDNA diluted 1/1000. The cycling conditions were 2 min at 94°C, 40 cycles of 94°C for 1 min, 57°C for 1.5 min, and 72°C for 1min. The final extension step was for 7 min at 72°C. The PCR primers were:

CcCPI-2 (up) 5' GTGAAGCCATGGTTGAACTT 3'
CcCPI-2 (low) 5' GTAATGATACTCAAGCCAGA 3'.

The lanes are labeled as follows: - M, Markers; +P, diluted plasmid containing the CcCPI-2 gene; R, root; S, stem; L, young leaves; F, flowers. - SG (G), LG (G), Y (G) and Red (G) are grain isolated from small green, large green, yellow and red cherries, respectively. - SG (P), LG (P), Y (P) and Red (P) are pericarp tissue isolated from small green, large green, yellow and red cherries, respectively;

Figure 10 shows the optimal alignment of the complete protein encoded by CcCPI-3 cDNA with other homologous full-length cysteine proteinases available in the NCBI. Shaded blocks indicate identical amino acids. Accession numbers of the EMBL database and percentage identities are given in parentheses. *Citrus x paradisi* (AAG38521; 42.4% identity), *Actinidia deliciosa* (AAR92223; 44.4% identity), and *Arabidopsis thaliana* (AAM64661; 44% identity);

Figure 11 shows the optimal alignment of the complete protein encoded by CcCPI-4 cDNA with other homologous full-length cysteine proteinases available in the NCBI. Shaded blocks indicate identical amino acids. Accession numbers of the EMBL database and percentage identities are given in parentheses. *Citrus x paradisi* (AAG38521; 23.6% identity), and *Arabidopsis thaliana* (AAM64661; 20% identity);

Figure 12 shows RT-PCR analysis of the expression of CcCPI-4 gene in different tissues of *Coffea arabica* CCCA2 (A) and *Coffea robusta* FRT-32 (B). The PCR reactions were carried out using 10 µl of each cDNA diluted 1/100. The cycling conditions were 2 min at 94°C, 40 cycles of 94°C for 1 min, 60°C for 1.5 min, and 72°C x 1min. The final extension step was for 7 min at 72°C.

PCR primers were:

CcCPI-4 (up) 5' CTACGGTCGCAGCCAAATC 3'

CcCPI-4 (low) 5' ACAACTGCACCTTCAATGTAC 3'.

The lanes are labeled as follows: - M, Markers; +P, diluted plasmid containing the CcCPI-4 gene; R, root; S, stem; L, young leaves; F, flowers. - SG (G), LG (G), Y (G) and Red (G) are grain isolated from small green, large green, yellow and red cherries, respectively. - SG (P), LG (P), Y (P) and Red (P) are pericarp tissue isolated from small green, large green, yellow and red cherries, respectively;

Figure 13 shows a Northern blot analysis of aspartic proteinase 2 (CcAP2) gene in different tissues of *Coffea arabica*, in which the lanes are labelled R: root, S: stem, L: young leaves, F: flowers; SG(G) and (P), LG(G) and (P), Y(G) and (P) and Red(G) and (P) are for grain and for pericarp, respectively, from small green, large green, yellow and red cherries, and SG(G), LG(G), Y(G) and R(G) for pericarp from small green, large green, yellow and red cherries respectively. Five micrograms of total RNA was loaded in each lane. Panel A demonstrates the ethidium bromide staining of large ribosomal RNA prior to blotting as a loading control and panel B is an autoradiogram showing the appearance of the CcAP2 mRNA in the specific tissues tested;

Figure 14 shows the cDNA sequence and the deduced amino acid sequence of CcCP-4. Lowercase: 5' and 3', non-translated regions; Uppercase: Open reading frame; Bold character: amino acid sequence; *: stop codon;

Figure 15 shows the alignment of the full sequence of the protein encoded by CcCP-4 cDNA with other full-length cysteine proteinases available in the NCBI database. This was done using the CLUSTAL W program in the MegAlign software (Lasergene package, DNASTAR). Shaded blocks indicate identical amino acids. Accession numbers are given in parentheses. Dacus carrota (JC7787); Ricinus communis (AF050756); Vicia sativa (Z34895); Phaseolus vulgaris (X56753); Helianthus annuus (AB109188); Glycine max Cys1 (AB092555); Glycine max Cys2 (AB092557); Canavalia ensiformis (P49046); Oryza sativa (AB004648); Vigna mungo (P12412); Pisum sativum (AJ004985);

Figure 16 shows the full length cDNA sequence CcCP-4 KDDL and the partial cDNA sequence CcCP-4 (KDEL) were aligned using the program ClustalW in Megalign;

Figure 17 shows the complete open reading frame of CcCP-4 (KDDL) and the partial open reading frame of CcCP-4 (KDEL) were aligned using the program ClustalW in Megalign;

Figure 18 shows the DNA sequence chromatograms for PCR amplified genomic DNA encoding the KDEL/KDDL region of the CcCP-4 gene. Rob, indicates a robusta variety and Arab, indicates an arabica variety;

Figure 19 shows Northern blot analysis of the expression of the Cysteine proteinase CcCP-4 gene in different tissues of *Coffea arabica*. The lanes are labeled as follows: - R, root; S, stem; L, young leaves; F, flowers. - SG (G), LG (G), Y (G) and Red (G) are grain isolated from small green, large green, yellow and red cherries, respectively. - SG (P), LG (P), Y (P) and Red (P) are pericarp tissue isolated from small green, large green, yellow and red cherries, respectively. Five micrograms of total RNA was loaded in each lane. Panel A demonstrates the ethidium bromide staining of the large ribosomal RNA prior to blotting as a loading control, Panel B is an autoradiogram showing the appearance of the CcCP-3 mRNA in the specific tissues tested;

Figure 20 shows RT-PCR analysis of the expression of CcCP-4 in the whole grain during germination. Sampling times were 0, immediately after sterilization treatment; 2D, 2 days after treatment; 3D, 3 days after treatment; 5D, 5 days after treatment; 1M, one month after treatment, 2M, two months after treatment; -, no DNA control; +P, diluted CcCP-4 plasmid DNA; M, molecular weight markers;

Figure 21 shows optimal alignment of the complete protein encoded by CcAP-1 cDNA with other homologous full-length aspartic proteinase sequences available in the NCBI. Shaded blocks indicate identical amino acids. Database accession numbers are given in parentheses. *Arabidopsis thaliana* (AY099617) and *Arabidopsis thaliana* (BAB09366); and

Figure 22 shows optimal alignment of the complete protein encoded by CcAP-2 cDNA with other homologous full-length aspartic proteinase sequences available in the NCBI. Shaded blocks indicate identical amino acids. Database accession numbers are given in parentheses. *Glycine max (*BAB64296), *Ipomoea batatas* (AAK48494), *Lycopersicon esculentum* (S71591) and *Nepenthes alata* (BAB20972).

Detailed Description of the Invention

As used herein, a "polynucleotide" is a nucleotide sequence such as a nucleic acid fragment. A polynucleotide may be a polymer of RNA or DNA that is single- or double-stranded, that optionally contains synthetic, non-natural or altered nucleotide bases. A polynucleotide in the form of a polymer of DNA may comprise one or more segments of cDNA, genomic DNA, synthetic DNA or mixtures thereof.

Similar nucleic acid fragments are characterised, in the present invention, by the percent identity of the amino acid sequences that they encode, to the amino acid sequences disclosed herein, as determined by algorithms commonly used by those skilled in the art. Suitable nucleic acid fragments (or isolated polynucleotides of the first to third aspects of the present invention) encode polypeptides that are at least 70% identical, preferably at least 80% identical, to the amino acid sequences disclosed herein. Preferred nucleic acid fragments encode amino acid sequences that are at least 85% identical to the amino acid sequences disclosed herein. More preferred nucleic acid fragments encode amino acid sequences that are at least 90% identical to the amino acid sequences disclosed herein. Still more preferred are nucleic acid fragments that encode amino acid sequences that are at least 95% identical to the amino acid sequences disclosed herein. Multiple alignment of sequences should be performed using the ClustalW method of alignment (Thompson *et al*, 1994, Nucleic Acids Research, Vol 22, p4673-4680; Higgins & Sharp 1989 Cabios. 5:151-153).

As used herein, the term "similar nucleic acid fragments" refers to polynucleotide sequences in which changes in one or more nucleotide bases result in substitution of one or more amino acids, but which changes either do not affect the function of the polypeptide encoded by the nucleotide sequence or do not affect the ability of nucleic acid fragment to mediate gene expression by gene silencing via, for example, antisense or co-expression technology. The term "similar nucleic acid fragments" also refers to modified polynucleotide sequences, in which one or more nucleotide bases is / are deleted or inserted, provided that the modifications either do not affect the function of the polypeptide encoded by the nucleotide sequence or do not affect the ability of nucleic acid fragment to mediate gene expression by gene silencing. It will, therefore, be understood that the scope of the present invention extends beyond the polynucleotide and polypeptide sequences specifically disclosed herein.

Similar nucleic acid fragments may be selected by screening nucleic acid fragments in the form of subfragments or modified nucleic acid fragments, for their ability to affect the level of the polypeptide encoded by the unmodified nucleic acid fragments in the plant or plant cell.

The term “operably linked” refers to the association of two or more nucleic acid fragments on a single nucleic acid fragment so that the function of one is affected by the other. “Regulatory sequences” refer to nucleotide sequences located upstream, within, or downstream, of a coding sequence and which influence transcription, RNA processing or stability, or translation of the coding sequence associated therewith. Regulatory sequences may include promoters, translation leader sequences, introns, transcription termination sequences and polyadenylation recognition sequences. When a regulatory sequence in the form of a promoter is operably linked to a coding sequence, the regulatory sequence is capable of affecting the expression of the coding sequence. Coding sequences can be operably linked to regulatory sequences in sense or antisense orientation.

The term “expression” refers to the transcription, and stable accumulation, of sense RNA (mRNA) or antisense RNA derived from the nucleic acid fragments of the present invention. Expression may also refer to the translation of mRNA into a polypeptide. Overexpression refers to the production of a gene product in a transgenic cell, that exceeds the level of production in normal, or non-transformed, cells. ”Altered levels” refers to the production of gene product(s) in a transgenic cell in amounts or proportions that differ from that of normal, or non-transformed, cells.

“Transformation” refers to the transfer of a nucleic acid fragment into the genome of a host cell, resulting in genetically stable inheritance. Host cells containing the transformed nucleic acid fragments are referred to herein as “transgenic cells”.

Standard recombinant DNA and molecular cloning techniques as used herein are well known in the art and are described more fully in Sambrook *et al* ”Molecular Cloning: A Laboratory Manual”; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 1989, which is incorporated herein by reference.

Examples

The following Examples illustrate the invention without limiting the invention to the same. In the examples, all parts and percentages are by weight and degrees are in Celsius, unless this is otherwise specified.

In the following Examples, these abbreviations have been used:
PCR : Polymerase chain reaction
RACE : Rapid amplification cDNA ends

From the above discussion and the Examples below, those skilled in the art can ascertain the essential features of the present invention, and without departing from the scope thereof can make various changes and modifications thereto, to adapt it to various usages and conditions as desired.

Production of cDNA libraries and screening

Production of Seed Specific RNA

Coffee cherries of the *Robusta* variety Q121 were harvested 30 WAF (weeks after flowering) at the ICCRI, Indonesia. The pericarps of these cherries were then removed and the remaining perisperm/endosperm material was frozen and ground to a powder in liquid nitrogen. The RNA was extracted from the frozen powder material using the method described previously for the RNA extraction of cacao seeds (Guilloteau, M. *et al*, 2003, Oil bodies in *Theobroma cacao* seeds: cloning and characterisation of cDNA encoding the 15.8 and 16.9 kDa oleosins. Plant Science Vol 164, 597-606). Poly A^+ RNA was prepared from approximately 250μg total RNA using the "PolyA Purist^{TM}" kit of AMBION (manufactured by Ambion, Inc.) according to their kit instructions.

Production of First Set of Seed cDNA clones

Approximately 50-100ng of this poly A^+ RNA was then employed in the synthesis of the first strand cDNA using "SuperScript^{TM} II RNase H^- reverse transcriptase (GIBCOBRL^{TM})and the SMART^{TM} PCR cDNA synthesis kit (Clontech) as follows. A reaction containing 2μl of 30 WAF poly A^+ RNA, 1 μL CDS oligo (SMART^{TM} PCR cDNA kit, Clontech), 1 μL Smart

II oligo (SMART™ PCR cDNA kit, Clontech), and 8 µL deionised H_2O. This mixture was heated to 72°C for 5 minutes and then placed on ice. Then the following was added; 1 µL 10 mM dNTPs, 4 µL SuperScriptII™ 1st stand buffer and 2 µL DTT. This mixture was put at 42°C for 2 minutes then 1 µL of SuperScriptII™ RNaseH$^-$ reverse transcriptase (200 units/µL GIBCO BRL™) was added and the mixture was incubated in an air circulating incubator at 42°C for a further 50 minutes.

After the reverse transcription reaction, the following PCR reaction was carried out. 98 µL of the Master Mix described in the SMART™ PCR cDNA kit (Clontech) containing Advantage™ 2 polymerase (Advantage™ 2 PCR kit, ClonTech) was set up on ice and then 3 µL of the 1st strand cDNA synthesis reaction described above was added. This 100 µL PCR reaction was then placed in a MJ Research PTC-150 HB apparatus and the following PCR conditions were run: 95°C for 1 minute, then 16 cycles of 95°C for 15 seconds, 65°C for 30 seconds, 68°C for 6 minutes. The amplified DNA was purified using the Strataprep™ PCR Purification Kit (Stratagene) according to the suppliers' instructions. The DNA, which was eluted in 50 µL deionized water, was then "polished" using the Pfu-1 polymerase reagents contained in the PCR-Script™ Amp cloning kit (Stratagene) as follows; 50 µL DNA, 5 µL 10 mM dNTPs, 6.5 µL 10 x Pfu-1 polishing buffer, 5 µL cloned Pfu-1 DNA polymerase (0.5 U/µl). This reaction was then incubated at 72°C for 30 minutes in a PCR apparatus with a heated cover (Perkin Elmer). Using the protocol described in the pPCR-Script™ Amp kit (Stratagene), the polished (blunted) PCR products were ligated into the Srf-1 digested pPCR-Script™ Amp SK(+) vector in the presence of Srf-1 enzyme and the ligation reaction products were transformed into the XL-10 Gold™ Kan ultracompetent E. coli cells. Selection for transformation with plasmids containing inserts was done using LB-Amp plates and IPTG and Xgal spread on the surface as described in the pPCR-Script™ Amp kit. White colonies were selected and the clones were named Dav1-1 etc.

Production of Second Set of seed cDNA clones with Size Selected cDNA

Seeds highly express a small number of proteins, such as the seed storage proteins (White *et al*, 2000, Plant Physiology, Vol 124, 1582-1594). When cDNA is prepared from such tissue, the very high level of the storage proteins and other seed specific proteins leads to a high level of cDNA "redundancy", that is, the population of cDNA produced contains high proportions of the same cDNA. In order to reduce the redundancy of cDNA made from coffee seed mRNA, and to selectively characterise long and weakly expressed cDNA, a second cDNA cloning strategy was also used. Using the products of the reverse transcriptase reaction described above, the following PCR reactions was set up using the Advantage^{TM} 2 PCR kit (ClonTech): 3 µL of the reverse transcriptase reaction, 5 µL 10 x Advantage^{TM} 2 PCR buffer, 1 µL dNTP's (10 mM each), 2 µL PCR primer (SMART^{TM} PCR cDNA kit, Clontech), 39 µL deionised water, and 1 µL 50 x Advantage^{TM} 2 polymerase mix. This PCR reaction was then placed in a MJ Research PTC-150 HB apparatus and the following PCR conditions were run: 95°C for 1 minute, then 16 cycles of 95°C for 15 seconds, 65°C for 30 seconds, 68°C for 6 minutes. At the end of the PCR, 1 µL 10% SDS was added with gel loading buffer, the sample was heated to 37°C for ten minutes. The sample was then split for loading onto a 0.7% agarose gel <u>without</u> ethidium bromide: 10% was loaded into a small well beside a DNA marker lane and the other 90% was loaded into a neighbouring large, preparation scale well. After the gel was run, the gel section with the size markers, plus the 10% reaction sample, were stained with ethidium bromide. This stained gel section was then used as a template to generate gel slices containing PCR amplified cDNA of different sizes from the cDNA present in the remaining unstained (preparation) part of the gel. Six gel slices were generated having the indicated size range of PCR fragments; A1A (0.8-1kb), A1B (1-1.5 kb), A2 (1.5-2.25 kb), A3 (2.25-3.25), A4 3.25-4 kb), and A5 (4-6.5 kb).

The DNA in each gel slice was eluted from the agarose using the QIAEX II kit from Qiagen following the suppliers instructions (for samples 3A, 4A, and 5A were heated for 10 minutes at 50°C and 1A, 1B, and 2A were heated for 10 minutes at room temperature). The purified double stranded cDNA was then re-amplified further by PCR with a TAQ enzyme mix which makes fragments having a 3' T overhang as follows: 30 µL of the gel isolated double stranded cDNA, 5 µL 10 x TAQ buffer (supplied with TAQ PLUS precision polymerase mix, Stratagene), 1 µL 40 mM dNTP's (each 10 mM), 2 µL PCR primer (SMART^{TM} PCR cDNA kit, Clontech), 0.5 µL TAQ PLUS precision polymerase mix (Stratagene) and 11.5 µL deionised water. The PCR reaction conditions were as follows: 95°C for 1 minute then 7 cycles 95°C for 15 seconds, 65°C for 1 minute, 72°C for 8 minutes, then 1 cycle at 95°C for 15 seconds, 65°C for 1 minute, 72°C for 10 minutes.

The PCR amplified DNA produced was then ligated into the vector pCR^{TM}-TOPO^{TM} and cloned into TOP10 *E. coli* cells using the TOPO^{TM} TA kit (Invitrogen) as described by the supplier. The clones were named by their order of isolation and their position in the sizing gel (for example, A2-1, A2-2, etc.).

Seed cDNA Screening and Preliminary Identification

The first set of white colonies obtained in Dav-1 library were screened by first determining the size of each insert by PCR amplifying the insert using the primers T3 and T7 which flank the cloning site used and examining the PCR amplified fragments on a gel.

Each white colony was resuspended in 200 µl sterile water and 10-30 µl of this was added to 5 µl 10X Taq polymerase buffer (Stratagene), 1 µl 10 mM dNTP mix, 2.5 µl 20 µM T3 primer, 2.5 µl 20 µM T7 primer, 1 µl DMSO, 0.5 µl Taq polymerase (Stratagene), and H_20 up to 50 µl final volume). The PCR reaction program used was 94°C for 1 min, then 30 cycles of 94°C for 1 min, 55°C for 1.5 min and 3.5 min at 72°C, and a final cycle of 7 min at 72°C. To reduce redundancy, the PCR inserts of similar size were subjected to digestion by the restriction enzyme Hae III.. Those PCR fragments with the same Hae III restriction pattern were not studied further. The plasmids of clones with PCR fragments >500 bp and which had unique Hae III restriction patterns were then purified by using the Qiawall 8 ultra plasmid kit (Qiagen) for 5' end dideoxy sequencing using the appropriate T7 or T3 sequencing primers coded in the flanking vector sequences. Because the inserts were not cloned in a directed fashion, it was first necessary to determine the 5' end of each clone by a Sca1 digestion of the purified plasmid DNA (the CDS SMART primer contains a Sca 1 site allowing the orientation of the insert to be determined). The DNA sequence data obtained was subsequently blasted against the non-redundant database protein in GENEBANK to obtain a preliminary annotation of each cDNA clone using the program BLASTX^{TM}.

Seed cDNA banks have a high level of redundancy. That is, a small number of seed mRNA have an unusually high level of expression, such as those encoding the seed storage proteins, and therefore their cDNA are very abundant in seed cDNA banks (White *et al*, 2000, Plant Physiology, Vol 124, 1582-1594). Therefore, as soon as the main redundant cDNA's were identified in the first round of sequencing the coffee seed cDNA, a pre-screening step was added for the white insert containing colonies prior to the determination of insert size. Four sequences were very highly expressed and the following specific primers sets were made for each of these redundant sequences,

1) 2S protein, contig 8A 5' AGCAACTGCAGCAAGGTGGAG 3' and contig 8B 5' CGATTTGGCACTGCTGTGGTTC 3' (55°C used in PCR, 114 bp fragment),
2) 2S protein contig 15A 5' GCCCGTGCTCCTGAACCA 3' and contig 15B 5' GTATGGTTGCGGTGGCTGAA 3' (55°C used in PCR,256 bp fragment),
3) Oleosin 15.5 contig 30A 5' ACCCCGCTTTTCGTTAT 3' and contig 30B TCTGGCTACATCTTGAGTTCT 3' (55°C used in PCR, 261 bp fragment), and
4) 11S protein contig 37A 5' GTTTCCAGACCGCCATCAG 3' and contig 37B 5' ATATCCATCCTCTTCCAACACC 3' (59°C used in PCR, 261 bp fragment).

The PCR reactions for this prescreen step were run as follows: 10-30 µl of the white colony in sterile H_2O, 5 µl 10X Taq buffer (Stratagene), 1 µl 10 mM dNTP, 2.5 µl of each primer at 20 µM, 1 µl DMSO, 0.5 µl Taq polymerase (Stratagene 10U/µl) and sterile H_2O was added to produce a final reaction total volume of 50 µl. The PCR program was 1 min at 94°C, then 30 cycles of 1 min at 94°C, 1.5 min at specific temperature for each primer pair, 2.5 min at 72°C, followed by 7 min at 72°C.

Full Length cDNA Insert Sequencing and Sequence Analysis

cDNA clones whose partial sequences showed initial homologies to proteinases and proteinase inhibitors were fully sequenced on both strands using the standard dideoxy primer walking strategy. The sequences are shown under SEQ ID Nos. 1, 3, 5, 7, 9, 11, 13 and 15. The full length sequences obtained were again blasted against the GenBank non redundant protein database using BLASTX to reinforce the preliminary annotation.

Sequence identities of sequence pairs were calculated using the ClustalW^{TM} program contained in the MegAlign^{TM} module of the Lasergene^{TM} software package (DNASTAR Inc). The default parameters were chosen as follows: (1- MULTIPLE ALIGNMENT PARAMETERS - Gap penalty 15.00, Gap length penalty 6.66, Delay divergent Seqs (%) 30, DNA transition weight 0.5, Protein Weight Matrix-Gonnet Series, DNA Weight Matrix IUB. 2- PAIRWISE ALIGNMENT PARAMETERS-Slow/Accurate (Gap Penalty 15.00, Gap Length Penalty 6.66), Protein Weight Matrix-Gonnet 250, DNA Weight Matrix-IUB) and the sequences used were either the full length nucleotide sequence of each cDNA or the full ORF (open reading frame) of each cDNA.

TABLE 2: Identity values between the nucleic acid and amino acid sequences of CcCP-1, CcCPI-1, CcAP-1 and CcAP-2 and related genes found in the non-redundant protein database of GenBank and those of WO 02/04617.

cDNA Sequences	nucleotide identity (%)	protein identity (%) (ORF)
CcAP1 vrs TcAP1	2.9	13.3
CcAP1 vrs TcAP2	2.4	9.8
CcAP2 vrs TcAP1	55.0	61.5
CcAP2 vrs TcAP2	55.1	61.3
CcCP-1 vs *Arabidopsis thaliana* putative cysteine proteinase (AY070063)	51.8	64.3
CcCP-1 vs Glycine max cysteine endopeptidase (Z32795)	49.1	61.3
CcCP-1 vs *Vicia sativa* cysteine proteinase precursor (Z99172)	49.0	60.9
CcAP2 vs *Lycopersicon esculentum* aspartic proteinase precursor (L46681)	65.9	71.1
CcAP2 vs *Ipomoea batatas* putative aspartic proteinase mRNA (AF259982)	71.7	69.6
CcAP2 vs *Nepenthes alata* NaAP4 mRNA for aspartic proteinase 4 (AB045894)	58.4	66.5
CcCPI-1 vs *Malus x domestica* cystatin (AY176584)	38.8	45.5

5' RACE PCR

The cDNA insert of clone A5-812 was found to contain introns. Therefore, to confirm the coding sequence of this protein, it was necessary to isolate a new cDNA containing the complete coding sequence. This was accomplished by using the SMART™ RACE cDNA amplification Kit (Clontech). The first strand cDNA used for the 5' RACE was made as already described for the cDNA libraries above. A gene specific primer rAP2 (5' CATATAATATTAAAAGCACCACCCATAA 3') was designed - this sequence is situated 92 pb from the poly (A) tail of A5-812 clone. This specific primer was then used with the Universal Primer Mix (UPM) in the CLONTECH kit in a PCR reaction under the following conditions; 2.5 µl of first strand cDNA product, 5 µl of 10X Advantage 2 PCR Buffer (CLONTECH), 1 µl of dNTP Mix (10 mM), 1 µl of 50X Advantage 2 Polymerase Mix (CLONTECH), 5 µl of "Universal Primer A Mix" (10X) (CLONTECH), 1 µl of rAP2 (10 µM) and sterile water was added to a final volume of 50 µl. PCR cycling conditions were 20 cycles of 30 sec at 94°C, 30 sec at 68°C and 3 min at 72°C, followed by a final extension reaction for 5 min at 72°C. A fragment of about 1700 pb was obtained, excised from the gel using "CONCERT™ Rapid Gel Extraction kit" (GibcoBRL). The isolated fragment was cloned in the pCR 4-TOPO vector and transformed into *Escherichia coli* using the Topo-TA cloning kit (Invitrogen). The plasmid obtained was then purified using a plasmid extraction kit (QIAfilter Plasmid Midi Kit, Qiagen, France) and the insert of this plasmid was double strand sequenced.

The DNA of clone A5-442 (AP1) was found to lack the 5' region of the cDNA. To isolate this region a 5' RACE was performed using the SMART™ RACE cDNA amplification Kit (Clontech). A sequence specific primer rAP1 (5'-TGGAGTCACAAGATGTCTCGACGAACTG-3') situated at 396 pb from the poly (A) tail was designed. This specific primer was then used with the Universal Primer Mix (UPM) in the CLONTECH kit in a PCR reaction under the following conditions; 2.5 µl of first strand cDNA, 5 µl of 10X Advantage 2 PCR Buffer (CLONTECH), 1 µl of dNTP Mix (10 mM), 1 µl of 50X Advantage 2 Polymerase Mix (CLONTECH), 5 µl of "Universal Primer A Mix" (10X) (CLONTECH), 1 µl of rAP1, and sterile water was added to a final volume of 50 µl. PCR cycling conditions were 20 cycles of 30 sec at 94°C, 30 sec at 68°C and 3 min at 72°C, followed by a final extension reaction for 5 min at 72°C. A fragment of about 2,000 bp was obtained, excised from the gel using "CONCERT™ Rapid Gel Extraction kit" (GibcoBRL). The isolated fragment was cloned in the pCR 4-TOPO vector and transformed into *Escherichia coli* using the Topo-TA cloning kit (Invitrogen). The plasmid obtained was then purified using a plasmid extraction kit (QIAfilter Plasmid Midi Kit, Qiagen, France) and the insert of this plasmid was double strand sequenced.

RNA preparation for Large Est Libraries:

RNA was isolated from dissected grain and pericarp tissues at various developmental stages, and from young leaves using the method described earlier. The varieties and tissues used to prepare the RNA to generate the different Est libraries were as follows: (1) young leaves, one variety (FRT-32); (2) pericarp (8 different developmental stages) from 5 varieties (FRT 32, FRT-31, FRT-400, FRT-4001, and Q121); (3) whole cherry, 22 weeks after fertilisation (WAF) from one variety (FRT-31); (4) grain, 18 + 22 WAF from five varieties (FRT 32, FRT-31, FRT-400, FRT-4001, and Q121); (5) grain, 30 WAF from 5 varieties (FRT 32, FRT-31, FRT-400, FRT-4001, and Q121) ; (6) grain, 42 WAF from five varieties (FRT 32, FRT-31, FRT-400, FRT-4001, and Q121) and (7) grain, 46 WAF from 2 varieites (FRT-32 and Q 121).

Production of cDNA clones, and DNA sequence analysis .

The cDNA clones for the various Est libraries were prepared as follows: Poly A$^+$ mRNA was isolated using the PolyATrack^{TM} mRNA Isolation System (System IV, Promega) according to the manufacturer's instructions for small scale isolation. The purified poly A$^+$ mRNA was then used to prepare cDNA for unidirectional cloning into the lambda phage as described in the ZAP-cDNA^{TM} library construction kit (cat # 200450 Stratagene). The mass excision protocol was to excise the pBlueScript phagemid from the Uni-ZAP XR vector and white colonies were obtained after plating on 150 mm LB-ampicilin agar plates with 80 ul x-gal (20mg/ml) and 16 ul IPTG (0.5M). Single colonies were randomly chosen to produce plasmid DNA which was then used for sequencing the 5' ends of the cDNA inserts.

The DNA sequences obtained produced an EST sequence (Expressed Sequence Tag) for each clone. All the Est sequence data from the 7 libraries was then clustered "*in-silico*", producing a unique group of sequences called the "unigene" sequence set. Thus, each "unigene" sequence theoretically corresponds to a distinct gene product. However, it should be noted that, because many unigenes only represent partial cDNA sequences, it is likely that some genes may be represented by two or more unigenes. Then a preliminary annotation of the unigene set was carried out with an automatic BLAST search where each unigene sequence was searched against the non-redundant GenBank protein database. This BLAST search approach produced the five best BLAST "hits" ("hits" with the lowest e-values) which is referred to as the "unigene annotation".

Northern-Blot Analysis

Freshly harvested roots, young leaves, stem, flowers and fruit at different stages of development (small green fruit (SG), large green fruit (LG), Yellow fruit (Y) and red fruit (R)) were harvested from *Coffea arabica* CCCA2 grown under greenhouse conditions (25°C, 70 RH) in Tours, France, and from *Coffea canephora* FRT32 grown either in Equador or ICCRI, Indonesia. The fresh tissues were frozen immediately in liquid nitrogen and total RNA was isolated from each tissue using the extraction procedure described above. A total of 5 µg of RNA was run on a 1.2% (w/v) denaturing RNA gel containing formaldehyde. The total RNA samples from each plant tissue were heated at 65°C for 15 min in presence of 7 µL "RNA Sample Loading Buffer" (without ethidium bromide, Sigma), and then put immediately on ice for 2 minutes before being loaded onto the 1.2% RNA gel. The gels were run at 60 Volts for 5 hours. The gel was then soaked twice in 10× SSC for 20 min. The RNA in the gel was transferred overnight by capillary transfer to a "Positive TM Membrane" (Qbiogene) in 10× SSC and the RNA was fixed by heating the blot for 30min at 80°C. Probes were generated using "Rediprime™ II random prime labelling system" kit (Amersham) in the presence of (P^{32}) dCTP. Hybridisation was carried out at 65°C for 24 h in hybridisation solution (5X SSC, 40µg/ml Denatured Salmon Sperm DNA, 5% [w/v] SDS, and 5× Denhardt's solution). Then, the membrane was washed twice at 65°C using 2X SSC, 0.1% SDS [w/v] and 1X SSC, 0.1% SDS [w/v] during 30 minutes each.

The Northern blot analysis shown in Figure 1 demonstrates that the coffee cysteine proteinase gene CcCP-1 gene is expressed in the *C. arabica* coffee cherry at all the stages tested, with yellow cherries exhibiting slightly higher levels of expression than the other stages. No expression was detected for this gene in the root, stem or leaves of *C. arabica*. Figure 2 shows another Northern Blot experiment examining the expression of CcCP-1 in *C. arabica* using a new preparation of RNA. For this experiment, the cherries for the four stages were dissected to generate pericarp tissue and grain tissue for each stage of cherry development. Total RNA was then extracted from these tissues. The results obtained show the same temporal pattern of expression for CcCP-1 during cherry development, but this new experiment additionally shows that CcCP-1 is primarily expressed at high levels only in the grain tissue of the cherries. No significant expression of the CcCP-1 gene is seen in the coffee cherry pericarp. This latter result supports the role of this gene product in the exclusive alteration of the protein, peptide and amino acid profile of the coffee grain under normal growing conditions.

We have generated EST libraries from coffee leaves, as well as from seed and pericarp tissues that have been dissected from different stages of developing coffee cherries. The detection of CcCP-1 ESTs in the different libraries (shown below – see Table 3) also demonstrates that this gene is expressed strongly in the grain, but is not expressed significantly in the pericarp or in leaves. The expression pattern of CcCP-1 during seed development is similar to that seen for its proposed homologous sequence of *Vicia sativa*, (CPR4 gene: Fischer, J. et al 2000. Plant Molecular Biology, 43, 83-101). These authors showed that CPR4 is not detected by Northern blotting in leaves, roots, or stem, further strengthening the argument that the CcCP-1 is grain specific. Altering the expression of CcCP-1 specifically in the grain as suggested here, such as by using a grain specific promoter for an antisense construct of CcCP-1 or an over-expression construct of CcCP-1, would not be expected to interfere with the metabolism in other tissues.

TABLE 3

Gene Name	Number of ESTs						
	Seed 18w	Whole Cherries 22w	Seed 30w	Seed 42w	Seed 46w	Pericarp	Leaf
CcCP-1	0	0	4	0	15	0	0

Optional alignment for CcCP-1 (Figure 2A) shows that this cDNA encodes a cysteine proteinase.

The Northern blot analysis shown in Figure 3 demonstrates that the coffee cysteine proteinase inhibitor gene CcCPI-1 gene is expressed in the *C. arabica* coffee cherry at all stages tested. However, in contrast to the expression seen for the cysteine proteinase CcCP-1, CcCPI-1 exhibits higher expression in the two early stages of coffee cherry development (small green and large green), and this gene is expressed at lower levels in the two later stages of cherry development. This expression pattern is consistent with the present hypothesis that the cysteine proteinase inhibitor protein (CcCPI-1) controls the activity level of a cysteine proteinase that is specifically expressed in seeds, such as CcCP-1, in the coffee cherry. A controlling protein such as the cysteine proteinase inhibitor protein can be expected to be expressed earlier than its target protein if it is necessary to control the level of activity of its target protein continuously from the time that the target protein is expressed. No expression was detected for this gene in the root, stem or leaves of *C. arabica*. It is noted that the similarity of the expression patterns for CcCP-1 and CcCPI-1 are consistent with the present hypothesis that these proteins could interact functionally.

The Northern blotting results (Figure 3) indicated that CcCPI-1 is expressed at all stages in the coffee cherry. However, this experiment did not determine whether the expression was in the whole cherry, or only in the pericarp or grain. Expression in the leaf was also not tested. However, the expression of CcCPI-1 in the different Est libraries (shown in Table 4 below) demonstrates that this gene is expressed specifically only in the grain, no expression was detected in the pericarp or leaves. This result further suggests that CcCPI-1 controls the activity level of a cysteine proteinase that is specifically expressed in seeds such as CcCP-1.

TABLE 4

Gene Name	Number of ESTs						
	Seed 18w	Whole Cherries 22w	Seed 30w	Seed 42w	Seed 46w	Pericarp	Leaf
CcCPI-1	0	0	1	0		0	0

The Northern blot analysis shown in Figure 4 demonstrates that the coffee cysteine proteinase inhibitor gene CcCPI-1 gene is expressed differently in the cherries of *C. canephora* (robusta) versus the cherries of *C. arabica*. First, the data of Figure 4 shows that the CcCPI-1 gene is expressed slightly earlier in *C. arabica*. Secondly, and more importantly, the CcCPI-1 gene is expressed in significantly higher levels in the *C.canephora* cherries. This difference in expression probably affects the level of the cysteine proteinase activity found in *C. arabica* versus *C. canephora* cherries. Because this class of protein is widely associated with insect resistance in plants, it is also likely that the high expression of the CcCPI-1 gene in C. canephora contributes to the higher disease resistance often seen for robusta varieties versus arabica varieties.

RT-PCR Analysis of CcCP-1 Expression During Grain Germination

To determine the expression of CcCP-1 during coffee grain germination, coffee fruit were harvested at the mature stage, rinsed with water, and the pericarp was taken off (each fruit normally contains two grains). The grain obtained were allowed to dry for one week in the open at room temperature. Before germination, the parchment and the silverskin (testa) of each grain were manually removed and grains were then sterilized byplacing in 1% (w/v) sodium hypochlorite for 1 hour, and then washed twice by sterilized, distilled water. For germination, 150 sterilized grains were placed individually in test tubes containing 10 ml of solid Heller growth medium H15, containing salts of Heller (Heller, 1953) and 7 g/ l agar and they were then incubated at 25°C, with 8 hours of light daily.

Three sets of ten grain were taken after 2 days, 3 days, 5 days, 1 month and 2 months of germination, and were immediately frozen in liquid nitrogen and stored at –80°C until RNA extraction. For the 1 and 2 month germination samples, the radicles associated with these samples were excised at sampling time and were frozen separately from the grain. Thirty sterilized grain were taken at T= 0 and frozen for use as a T(0) control.

4 µg of DNase-treated total RNA extracted from each sample was used to synthesize cDNA using hexamer oligonucleotides according to the protocol of the Superscript II Reverse Transcriptase (Invitrogen, Carlsbad, CA). A fragment of the coffee ribosomal protein L39 gene was amplified for each cDNA sample as a control for the cDNA synthesis step. The PCR reactions were performed using 50 µl reactions containing 10 µl of a 1/100 dilution of the cDNAs, 1µM each primer, 5 µl of 10X ThermoPol PCR buffer (10 mM $(NH_4)_2SO_4$, 2mM $MgSO_4$, 20 mM Tris-HCl, pH 8.8 at 25 °C, 10 mM KCl, and 0.1% Triton X-100) and 2.5 units of Taq polymerase (New England Biolabs, Beverly, MA). The cycling conditions were 2 min at 94°C, followed by 35 cycles of 94°C for 1 min, 60°C for 1.5 min, and 72°C x 2.5 min. The final extension step was for 7 min at 72°C. The following primers were used for amplification by PCR: CcCP-1 up 5' ACCGAGGAGGAGTTTGAGGCTACG 3'and CcCP-1 low: 5' ACGCTTCCCCCATGAGTTCTTGA 3', yielding cDNA products of 726 bp. The primers for the RPL39 protein were:

A5-1750-upper	5' **TGGCGAAGAAGCAGAGGCAGA** 3'
A5-1750-lower5'	5' **TTGAGGGGGAGGGTAAAAAG** 3'

RT-PCR was used to determine the expression of CcCP-1 during the different stages of germination. The results obtained demonstrate that CcCP-1 transcripts are detected in the whole grain at all the germination times tested (Figure 5). It has previously been shown by Fischer, J. et al 2000 (Plant Molecular Biology, 43, 83-101) that RNA of the proposed CcCP-1 homologue CPR4 from *V. sativa* is also expressed in both the embryo axis and the cotyledons of *V. sativa* seeds during germination.

Western Blot Analysis

The leaf and cherry tissues analysed were from *Coffea arabica* CCCA2, and prior to use, the tissues were stored frozen at –80°C. The grain and pericarp tissues of the cherries at different stages of development were dissected separately with as little thawing of the pericarp as possible. These different tissues were then rapidly ground to a fine powder, such as can be done using liquid nitrogen with a pre-frozen mortar and pestle. A protein extract was prepared from this tissue using a modified version of the extraction procedure described by Tanaka et al., 1986 (Plant Physiology, 81 802-806). The buffers used were:

Tanaka buffer:

• Sucrose	0.7 M
• Tris-HCl pH 8	0.5 M
• β-mercapto-ethanol	2% ($^{v}/_{v}$)
• NaCl	0.1 M

And just before using this buffer add:

• EDTA	5 mM
• PMSF	2 mM

Gel loading buffer:

• Glycerol	15% ($^{v}/_{v}$)
• β-mercapto-ethanol	2% ($^{v}/_{v}$)
• SDS	3% ($^{v}/_{v}$)
• Tris-HCl pH 6.8	62.5mM

A few hundred milligrams of the frozen ground powders were added to 650µl of Tanaka buffer. The proteins were extracted with the addition of one volume of Tris saturated phenol pH8 (ie. saturated with 10 mM Tris-HCl pH8). Each sample was mixed vigorously for 20 min and then centrifuged for 20 min at room temperature at 13 000 g. After centrifugation, the proteins are in the phenolic phase. 20ul samples were kept for analysis (see below) and the remaining proteins in the phenol phase were precipitated overnight at –20°C following the addition of five volumes of methanol containing 0.1 M ammonium acetate. Subsequently, the samples were centrifuged for 20 min at room temperature at 13 000 g, and the resulting pellets were washed two times in 500 µl of methanol containing 0.1 M ammonium acetate. The pellets obtained were resuspended in 30 µl of gel loading buffer until protein quantification.

The protein in 20µl samples of the phenolic phase were also precipitated as above, and the final pellet was resuspended in the sample buffer of the BioRad D_C Protein assay Kit. Quantification of total protein in this sample was carried out using the BioRad D_C Protein assay kit as described by the supplier. Subsequently, all the main samples were adjusted to give 5µg/µl by addition of gel loading buffer.

Samples containing approximately 50 ug protein of each sample were separated by electrophoresis in an SDS-polyacrylamide gel (12% tris-glycine, (Novex® Invitrogen™). The proteins were then transferred to a PVDF membrane by electroblotting using standard protocols. Non-specific binding sites on the membrane were blocked by incubating the membrane in 10% non-fat dried milk in TBS buffer (BioRad™), for one hour at room temperature or overnight at 4°C. The blotted proteins were probed for two hours at room temperature or overnight at 4°C with a polyclonal antibody (dilution 1/5000e in TBS 10% non-fat dried milk), raised against the predicted homologue of CPR4 from *Vicia sativa,* which was kindly donated by A. Schlereth and K. Müntz, Institut für Pflanzengenetik und kulturpflanzenforchung (IPK), Germany (A.Schlereth, C.Becker, C. Horstmann, J. Tiedmann and K.Müntz 2000, Journal of Experimental Botany, 51:1423-1433). The membrane was then washed 3 times for 20 minutes in TBS + 0.1% Tween 20 buffer. The membrane was subsequently incubated one hour with a secondary antibody labeled with horseradish peroxidase (Goat anti-rabbit Ig, Immunopure®, Pierce™). The membrane was then washed 2 times for 20 minutes in TBS + 0.1% Tween 20 buffer, then once for 20 minutes in TBS. The presence of the enzyme coupled to the second antibody was visualized by chemiluminescence detection using the enhanced ECL+® system (Amersham Life Science) as described by the supplier.

The results obtained show that a polypeptide of approximately 41kDa, which corresponds closely with the predicted molecular weight of the CcCP-1 precursor polypeptide (43 735 Da), is detected at all the stages of grain maturation tested, but is not detected in the pericarp tissue (Figure 6). This protein expression pattern is similar to that seen for the CcCP-1 mRNA (Figure 2). Another polypeptide of approximately 22 kDa is also detected in the grain at the yellow stage and red stage, but in smaller quantities than 41 kDa polypeptide. The size of this second polypeptide is consistent with the predicted size of the mature form of CcCP-1 (25, 239 Da). The predicted size of the mature CcCP-1 after processing was determined by a protein alignment between the complete ORF sequence of CcCP1 and the sequence of the predicted mature form of CPR4 (*Vicia sativa* –accession# Z99172, 60.9% identity with CcCP1). The N-terminal site of the CPR4 polypeptide processing to generate the mature form was predicted by sequence comparison with other papain-like CPR polypeptides (J.Fisher, C.Becker, S. Hillmer, C. Horstmann, B. Neubohn, A. Schlereth, V. Senyuk, A. Shutov and K. Müntz. 2000 Plant molecular biology 43:83-101). Interestingly, in contrast to the results presented here, where both the precursor and mature forms of CcCP-1 are detected during grain development, only the mature form of the CPR4 polypeptide was detected in developing seeds and also during the germination of *V. sativa* seeds (Fisher and al, 2000).

RT-PCR analysis of gene expression for robusta variety FRT-32.

Different tissues of FRT-32 were prepared and total RNA was extracted from these tissues by the method described earlier. cDNA was prepared from DNase-treated total RNA as described above for the RT-PCR experiments with arabica cDNA. Then specific PCR reactions were run using the reaction conditions described above for the RT-PCR experiments with arabica cDNA. The specific amplification conditions and oligonucleotide primers used given in the Figure legend for each experiment.

CcCPI-1

a) Optimal alignment for CcCPI-1 (Figure 6A) showing that this cDNA encodes a cysteine proteinase inhibitor.

b) RT-PCR expression data for CcCPI-1 (Figure 7) in arabica and robusta. The PCR reactions were performed as previously described and the cycling conditions and the PCR primers used are given in the Figure legend. These data compliment and extend that data presented earlier for the arabica expression in that it shows CcCPI-1 is only expressed in the grain and not pericarp. Weak expression of this gene was also detected in flowers, a result not seen previously by Northern blot analysis. The RT-PCR expression in robusta was also determined (Figure 7). It is the same expression pattern as seen for arabica except that no expression was detected in flowers or in the small green grain. The absence of expression seen for the small green stage of robusta is also seen for other genes and is thus not unique to the CcCPI-1 gene.

Table 5: Occurrence of Est's for Cysteine Proteinase Inhibitor genes CPI-2, CPI-3 and CPI-4 in different Est Libraries.

Cysteine proteinase Inhibitor	Number of ESTs						
	Seed 18w	Whole Cherries 22w	Seed 30w	Seed 42w	Seed 46w	Pericarp	Leaf
CcCPI-2	0	2	12	0	1	1	0
CcCPI-3	0	0	1	0	2	0	0
CcCPI-4	0	0	1	0	0	0	6

CcCPI-2

a) Optimal Alignment for CcCPI-2 (Figure 8) showing that this cDNA encodes a cysteine proteinase inhibitor.

b) RT-PCR expression data for CcCPI-2 (Figure 9) in arabica and robusta. The PCR reactions were performed as previously described and the PCR primers used are given in the Figure legend. These data show that CcCP-2 is expressed in all tissues and thus the protein product of this gene probably plays an important role in controlling one or more cysteine proteinases present in these tissues. The numbers of Est's in each library seen in Table 5 above suggest that CPI-2 may be expressed more in grain (seed) at 30 weeks after fertilisation than in leaves, pericarp, or seeds 46 weeks after fertilisation.

CcCPI-3

a) Optimal Alignment for CcCPI-3 (Figure 10) showing that this cDNA encodes a cysteine proteinase inhibitor.

b) No RT-PCR expression data is currently available for this cysteine proteinase inhibitor. However, the "*in silico*" expression of this gene, as determined by the number of Est's appearing in each library (Table 5 above), indicates that CcCPI-3 is expressed in coffee grain (present in seed libraries "Seed30w" and "Seed46w" i.e. 30 and 46 weeks). The absence of Est's for this gene in the pericarp, leaf or whole cherry suggests that this gene may be a grain specific gene.

CcCPI-4

a) Optimal Alignment for CcCPI-4 (Figure 11) showing that this cDNA encodes a cysteine proteinase inhibitor.

b) RT-PCR expression data for CcCPI-4 (Figure 12) in arabica and robusta. The PCR reactions were performed as previously described and the PCR primers used are given in the Figure legend. The data obtained show that this gene is significantly expressed, in arabica, in leaves, flowers and in grain at the red stage. Because close examination of the original gel (Panel A : arabica) indicates that there are also weak bands detected in the small green grain and large green pericarp lanes, this gene may also be weakly expressed, in arabica, in the grain and pericarp at all the stages of cherry development studied. The data obtained for robusta show that this gene is significantly expressed in leaves, flowers, small green grain and large green grain. Only one Est for CcCPI-4 was found in the seed or pericarp libraries (Table 5 above), indicating that expression of this gene in the grain and/or pericarp is relatively low or is confined to small defined regions of these two tissues.

In each case for the Cysteine Proteinase Inhibitor (CPI) genes, the over-expression or inhibition of the expression of these genes during grain development (that is, under the control of a very strong grain specific promoter such as the coffee 11S promoter) is expected to alter the protein peptide and amino acids profiles in the mature grain (and thus the level of flavour precurors).

Germination and RT-PCR Analysis

Sterilized, dried *C. arabica* CCCA2 grain (parchment and silverskin removed) were placed individually in test tubes containing 10 ml of solid Heller growth medium H15 and 7 g/ 1 agar and were incubated at 25°C, with 8 h of light daily. After 2 days, 3 days, 5 days, 1 month and 2 months of germination, three grains were taken, and when present, the radicles were removed and both grain and radicles were immediately frozen in liquid nitrogen and stored at –80°C until RNA extraction. Similarly dried and sterilized non-germinated grains (T0) were used as control. RNA was extracted from the grain samples as described earlier. DNase-treated total RNA extracted from each sample was used to synthesize cDNA using oligo $(dT)_{20}$ as a primer according to the protocol of the Superscript II Reverse Transcriptase kit (Invitrogen, Carlsbad, CA). A PCR reaction was then carried out using aliquots of each cDNA reaction. (50 μl reactions containing 10 μl of the 1/10 diluted cDNAs, 1μM each primer, 5 μl of 10X ThermoPol PCR buffer, 200um dNTPs and 2 units of Taq polymerase (New England Biolabs, Beverly, MA). The cycling conditions were 2 min at 94°C, 40 cycles of 94°C for 1 min, 54°C for 1.5 min, and 72°C x 2.5 min. The final extension step was for 7 min at 72°C. PCR primers were CP-4KDDL61 : 5'-
GAAGAACTCATGGGGAACAGGAT - 3'
CP-4KDDL345 : 5'- TTATTCAAACCATCACAGGAGCAG - 3'

Genomic PCR and DNA sequencing of the purified PCR fragments

Genomic DNA of five different coffee varieties (FRT-07, FRT-19, FRT-32, CCCA2, and GPFA57) was used in the PCR reaction described above for the germination RT-PCR expression study. PCR products of the expected size were obtained and these fragments were purified from the gel. The PCR amplified DNA was then subjected to a second round of PCR amplification and the DNA obtained from this sequencing reaction was then sequenced using the same primers as used for the ampification.

Isolation and Characterization of Cysteine Proteinase CcCP-4

Using a collection of Est's (Expressed Sequence Tags) made with RNA isolated from 1) coffee grain at different stages of development, coffee pericarp tissue at different stages of development, and from leaves, we have isolated a full length cDNA encoding a coffee cysteine proteinase which has a C-terminal KDDL sequence. We have named this cDNA CcCP-4 (KDDL) (Figure 14). The alignment of the protein encoded by this cDNA with other highly homologous plant cysteine proteinases is shown in Figure 15. This alignment data, and the related Blast searches, clearly show that the protein encoded by the coffee CcCP-4 (KDDL) sequence is a member of the plant KDEL containing cysteine proteinase family (Figure 15). The precise identities between CcCP-4 (KDDL) and the most homologous database sequences is given in Tables 6A and 6B.

Table 6A: Identity of the *Coffea canephora* cysteine proteinase CcCP-4 (KDDL) amino acid sequence with the amino sequences of the most homologous GenBank sequences

Coffea canephora cysteine proteinase	**Gene name (accesion number)**	**% identity protein**
CcCP-4 (KDDL)	Dacus carrota (JC7787	73
	Vigna mungo (P12412)	69
	Glycine max Cys1 (AB092555)	70
	Glycine max Cys2 (AB092557)	68
	Vicia sativa (Z34895)	64

Table 6B: Identity of the *Coffea canephora* cysteine proteinase CcCP-4 (KDDL) nucleic acid (cDNA) sequence with nucleic sequences of the most homologous GenBank sequences

Coffea canephora cysteine proteinase	**Gene name (accesion number)**	**% identity DNA**
CcCP-4 (KDDL)	Dacus carrota (JC7787)	55
	Vigna mungo (P12412)	61
	Glycine max Cys1 (AB092555)	49
	Glycine max Cys2 (AB092557)	62
	Vicia sativa (Z34895)	60

Obviously, the coffee CcCP-4 KDDL sequence obtained has one important difference from nearly all the other sequences shown in Figure 15, that is, it does not have the expected endoplasmic reticulum (ER) retention sequence (the C-terminal KDEL sequence) but a varient of this sequence, ie. KDDL. By testing the capabilities of variations in the C-terminal KDEL sequence to direct retention in the ER in plant cells, Denecke et al (Denecke, J., De Rycke, R., and Botterman, J. 1992 EMBO J. 11, 2345-2355) have previously shown that C-terminal variants such as SDEL, KDDL, KDEI and KDEV can produce a complete loss of endoplasmic reticulum retention function. Therefore, the presence of the KDDL sequence in the coffee homologue of the plant KDEL cysteine proteinase was unexpected. Table 7 shows that the unigene containing the cDNA CcCP-4 (KDDL) has 21 Est's. Therefore, we then examined the sequence of other Est's in this unigene and we found that seven of these Est's contained good sequence data for the KDDL region. Of these seven cDNA sequences, six had the KDDL sequence and one had a KDEL sequence. We subsequently isolated the cDNA clone with a KDEL C-terminal sequence and obtained the complete sequence for this partial cDNA clone. The DNA and protein sequences obtained are shown in Figures 16 and 17 respectively.

Table 7. Number of Est's in the unigene containing the full length cDNA CcCP-4 (KDDL).

Cysteine proteinase Name	Unigene	Number of ESTs						
		Seed 18w	Whole Cherries 22w	Seed 30w	Seed 42w	Seed 46w	Pericarp	Leaf
CcCP-4	125103	0	0	8	0	13	0	0

The cDNA encoding the sequence for CcCP-4 (KDEL) shown in Figure 16 is only a partial cDNA, that is, it is only 817 bp long versus 1336 for the full length cDNA clone CcCP-4 (KDDL)). The partial cDNA CcCP-4 (KDEL) has 8 single nucleotide residue changes from the equivalent sequence found in the cDNA clone CcCP-4 (KDDL), although only two of these nucleotide changes lead to a change in the amino acid sequence of the open reading frame (Figure 17). In the 3'untranslated region, there are 3 clear nucleotide changes. In addition, there is also an insertion of 12 nucleotides in 3' untranslated region of the CcCP-4 (KDEL) cDNA sequence that appears to be within a micro-satellite region. The data just presented uncovered two different and important molecular markers for these two alleles of the coffee CcCP-4 gene, one is an SNP associated with the functionally important KDEL site, and the other is a microsatellite marker associated with the 3' untranslated region of this gene. The latter point is important as microsatellite sequences are usually considered genetic markers with high variability and therefore it is likely that other alleles of this gene could be found using this microsatellite containing region.

In order to examine the distribution of the two alleles of the CcCP-4 gene identified above in different varieties of arabica and robusta, a small region of the genomic sequence harboring the CcCP-4 gene was amplified by PCR from five different genotypes. The PCR fragments of the expected size (207 base pairs) were obtained from each genomic DNA sample and these PCR products were gel purified and then re-amplified to generate sufficient DNA for direct DNA sequencing of the PCR product. The results obtained from the sequencing reactions are shown in Figure 18. The sequencing chromatograms for the five sequences show that the two arabica varieties tested clearly had KDEL sequences and the three robusta varieties examined had KDDL sequences. This result implies that the KDDL allele could be restricted to robusta varieties, and that it is not found in arabica varieties. While no KDEL sequence was found in the three robusta varieties studied here, the discovery of one KDEL sequence in the Cornell Est library indicates that this allele can exist in at least some some robusta clones.

The expression of the CcCP-4 gene was studied using Northern blot and RT-PCR analysis. Figure 19 shows the result obtained from the Northern blot experiment using RNA extracted from different developmental stages of coffee grain and pericarp, as well as RNA isolated from roots, young leaves, stems, and flowers of an arabica variety. The data obtained using a CcCP-4 (KDDL) probe, which has approximately 98% homology with the known DNA sequence of the CcCP-4 (KDEL) allele, showed that CcCP-4 is expressed only in the grain. No expression was detected in the pericarp, or in the roots, stem, flowers, and leaves. Due to the very high level of identity between the two alleles, the CcCP-4 (KDDL) probe is expected to hybridize to transcripts from both alleles. A similar experiment using RT-PCR analysis also showed the same expression profile for the CcCP-4 gene.

Expression of CcCP-4 was also studied in the whole seed during germination using RT-PCR analysis. This experiment used primers that are common to both the CcCP-4 KDEL and the CcCP-4 KDDL alleles. The results of this experiment are shown in Figure 20. CcCP-4 transcripts were detected at all the germination stages tested, although the level of transcripts appeared to dip slightly at 3 days and then begin to increase again as germination progressed (with the highest levels in 1 and 2 month samples).

Ling et al. (Ling, J.-Q., Kojima, T., Shiraiwa, M., and Takahara, H., 2003 Biochim. Biophys. Acta 1627, 129-139 have isolated two cDNA from soybean cotyledons encoding KDEL containing cysteine proteinases. These two cDNA had a 93.5% similarity at the DNA level, and were expressed in roots, flowers, and during seed development. No expression was detected by Northern blotting in developing or mature seeds, although expression was detected in mature pods. A cDNA encoding a KDEL containing cysteine proteinase was also isolated from carrot (Sakuta, C., Oda, A., Konishi, M., Yamakawa, S., Kamada, H., and Satoh, S. 2001 Biosci. Biotechnol. Biochem. **65**, 2243-2248.) Transcripts of this gene were detected in mature dry seeds, and in whole germinating seeds at day 2 and day 3 after imbibition. The expression of this gene in other carrot tissues or during seed development was not presented. Another KDEL containing protein, and its corresponding cDNA have been isolated from *V. sativa* (Fischer, J, Becker, C., Hillmer, S., Horstmann, C., Neubohn, B, Schlereth, A., Senyuk, V., Shutov, A., and Muntz, K. (2000) Plant Molecular Biol. **43**, 83-101.). Using Northern blotting, transcripts for this gene were detected in the cotyledons during germination, but not in the embryo axis of the germinating seeds. No transcripts were detected in maturing seeds, mature seeds, or in leaves and roots.

The results presented here show that the coffee KDEL type cysteine proteinase exhibits some novel, and unexpected features. First, we have discovered that robusta coffee grain expresses a KDEL type CP gene which has a single mutation in the sequence coding for the KDEL region resulting in change from KDEL to KDDL. Based on the data of Denecke et al. (1992), this particular alteration in the retention sequence is expected to alter the cellular localization and/or control of the robusta CcCP-4 (KDDL) protein. We propose that the presence of a transcribed copy of the CcCP-4 KDDL gene can produce a significant change in the peptide/amino acid profile in the coffee grain relative to varieties with the CcCP-4 KDEL sequence. We have also shown here that the KDEL type cysteine proteinase of coffee, while showing the expected expression during grain germination, is also unexpectedly expressed during all the grain development stages studied. As noted above, so far, there are no clear data in the published literature demonstrating a significant expression of a KDEL type cysteine proteinase during seed development in other plants, although its transcripts have been detected in mature carrot seeds (Sakuta et al., 2001).

The novel properties of the coffee KDEL type cysteine proteinase presented above probably have an important effect on the peptide and amino acid profiles in the mature grain of arabica and robusta, and therefore alter this pool of critical coffee flavour precursors. Considering that transcripts for the KDEL type cysteine proteinase are present in the mature grain, it is also possible that the KDEL type protein could be activated during the wet processing of coffee and thereby further alter the peptide/amino acid profile of wet process coffee grain. The work described has generated molecular markers (SNP's and a microsatellite marker) that can be used in classical selection and breeding work to obtain coffee varieties with specific alleles of the KDEL type cysteine proteinase gene (which will have concomitant alterations in protein/peptide/amino acid profiles). For example, varieties of robusta could be selected/bred which have only the CcCP-4 KDEL allele, or have only low expression levels of the CcCP-4 KDDL allele. Further, using genetic modification techniques it can be envisioned to alter the KDEL type cysteine proteinase activity in coffee, or in other plants, by the seed specific over-expression of the KDEL or KDDL type cysteine proteinases. Alternatively, the levels of the KDEL type cysteine proteinase can be reduced using antisense, sense or RNAi technologies. In both cases, the protein/peptide/amino acid pool in the resulting transformed plants will be altered, leading to new profiles of the protein/peptide/amino acid flavour precursor pools.

The Northern blot analysis shown in Figure 13 demonstrates that the coffee aspartic proteinase CcAP-2 gene is expressed in both the grain and the pericarp of the *C. arabica* coffee cherry at all cherry development stages tested. The CcAP-2 gene also has a relatively high expression in roots. When the film is exposed longer, CcAP-2 expression was also detected in the tissues of *C. arabica* stems, leaf, and flowers.

CcAP-1 and CcAP-2

Figures 21 and 22 show that each of CcAP-1 and CcAP-2 encode an aspartic proteinase.

Overexpression and under-expression of the CcCP-1 CcCP-4, CcAP-1 and CcAP-2 proteinase gene sequences and the CcCPI-1,2,3, and 4 proteinase inhibitors in coffee seeds. It is expected that the major storage protein profile and the amino acid/peptide profile can be changed in the mature coffee grain by altering, either up or down, the expression of one or more of the genes disclosed herein.

Methods for the overexpression of a gene of interest are well known in the art. Such methods consist of creating a chimeric gene of three major components, 1) a promoter sequence at the 5' end of the gene, preferably in the current application a seed specific promoter such as the coffee seed specific promoter described in Marraccini et al. 1999 (Marraccini *et al* 1999 Molecular cloning of the complete 11S seed storage protein gene of *Coffea arabica* and promoter analysis in transgenic tobacco plants, Plant Physiol. Biochem. Vol 37, 273-282, and WO 99/02688), 2) the entire coding sequence of the gene to be expressed, and 3) a 3' control region such as the 3' region from the nopaline synthase gene from the T-DNA of the Ti plasmid of *Agrobacterium tumefaciens*. Then, the chimeric gene can be cloned into an *Agrobacterium tumefaciens* transformation vector, and this vector can be transformed into an *Agrobacterium tumefaciens* strain for use in coffee transformation which has been described in detail by Leroy *et al* 2000, (Leroy *et al* 2000 Genetically modified coffee plants expressing the *Bacillus thuringiensis cry1Ac* gene for resistance to leaf minor. Plant Cell Reports 2000, 19, 382-389). Plants with stable transformation inserts can then be screened for those which overexpress the specific genes used in the transformation experiment specifically in mature seeds using methods such as detection of gene overexpression or protein activity overexpression versus seeds from mock transformed plants.

For example, a person well skilled in the art can produce a recombinant construct composed of 1) the longest coffee 11S gene promoter sequence described in Marraccini et al. (1999), 2) the full length cDNA sequence of CcCP-1, or of CcCP-4 (KDDL) without the poly A tail, and 3) a known transcription terminator sequence such as the well studied nopaline terminator. It is also possible that higher levels of over-expression for the recombinant constructs could result from the substitution of the 5' non-coding region of the CcCP4 or other cDNA sequences with the 5' non-coding region of the coffee 11S gene or the 5' non-coding regions of other strong seed specific promoters of either coffee or other related plant species. The recombinant gene sequences can then be inserted into an appropriate site of the *Agrobacterium* T-DNA vector described in Leroy et al. The T-DNA vector thus constructed can be put into an appropriate *Agrobacterium* strain, such as the strain described in Leroy et al., and the T-DNA containing *Agrobacterium* can be used to transform coffee following the method detailed in Leroy et al.

It is well known in the art that the expression of known gene sequences can be reduced or completely blocked by antisense suppression and by gene expression using nucleic acid fragments representing less than the entire coding region of a gene, and by nucleic acids that do not share 100% sequence identity with the gene to be suppressed. In this case, the sequences chosen for the particular antisense suppression or cosuppression experiment will replace the full length gene in the chimeric gene construction scheme presented above. The resulting antisense suppression or cosuppression chimeric constructions are again cloned into an *Agrobacterium tumefaciens* transformation vector, and transformed into *Agrobacterium tumefaciens* strain for use in coffee transformation as described above. Plants with stable transformation inserts can then be screened for those with reduced expression of the specific gene sequences used in the seeds of the transformed plants. The reduced expression can be detected by techniques such as Northern blotting; semi quantitative RT-PCR, and/or quantitative RT-PCR.

Another method for reducing, or eliminating, the expression of a gene in plants is to use the small portions of the gene sequences disclosed herein to produce RNA silencing via using RNAi (Hannon, G.J., 2002, Nature, Vol 418, 244-251; Tang *et al*, 2003, Genes Dev, Vol 17, 49-63). In this approach, small regions of one or more of the sequences disclosed herein are cloned into an *Agrobacterium tumefaciens* transformation vector as described above which has a seed specific promoter and an appropriate 3' regulatory region. This new inserted sequence for RNAi should be constructed so that the RNA produced forms an RNA structure in vivo which result in the production of small double stranded RNA in the transformed cells and whereby these small double stranded RNA sequences trigger the degradation of the homologous mRNA in these transformed cells.

Screening for naturally occurring variations in the CcCP-1, CcCP-4, CcAP-1, CcAP-2, CcCPI-1, CcCPI-2, CcCPI-3, CcCPI-4 genes and creating new mutations in these genes. The sequences disclosed herein can be used to screen natural populations for allelic variants in these genes. This can be accomplished by using the CcCP-1, CcCP-4, CcAP-1, CcAP-2, CcCPI-1, CcCPI-2, CcCPI-3 and CcCPI-4 sequences as probes in a search for naturally occurring RFLP's (restriction fragment length polymorphisms) in genomic DNA from different coffee plant varieties. A more powerful method to find allelic variants is to use the mutation screening technology associated with the TILLING method (Till, B.J., *et al* 2003 Large scale discovery of induced point mutations with high-thruput TILLING. Genome Research Vol 13, 524-530). In this case, once a specific gene sequence has been isolated and cloned, such as CcCP-1, CcCP-4, CcAP-1, CcAP-2, CcCPI-1, CcCPI-2, CcCPI-3 and CcCPI-4 sequences herein, the mutation screening technique associated with the TILLING method can be used to identify sequence variants between the cloned sequence and the corresponding cDNA or genomic sequence in different varieties. Using PCR primer pairs coding for DNA segments of 700-1100 base pairs, the known cloned gene can be scanned for naturally occurring sequence variations in different varieties. In the ideal situation, one or more sequence variants could also be correlated with a particular phenotypic variation thereby identifying a genetic marker for this phenotypic variant.

Additionally, using the sequences disclosed herein for CcCP-1, CcCP-4, CcAP-1, CcAP-2, CcCPI-1, CcCPI-2, CcCPI-3 and CcCPI-4, application of the full TILLING method can be used to create and detect new mutants in these genes and thus produce plants containing these specific mutants. For example, using the full TILLING method, coffee plants could be created which have specific mutations, such as a missense mutation in the coding sequence which inactivates the gene target of interest.